VAR(1) Based Models do not Always Outpredict AR(1) Models in Typical Psychological Applications

Kirsten Bulteel*, Merijn Mestdagh*, Francis Tuerlinckx, and Eva Ceulemans

KU Leuven

Author note

Kirsten Bulteel, Department of Psychology, KU Leuven; Merijn Mestdagh, Department of Psychology, KU Leuven ; Francis Tuerlinckx, Department of Psychology, KU Leuven; Eva Ceulemans, Department of Psychology, KU Leuven.

Kirsten Bulteel and Merijn Mestdagh are doctoral research fellows with the Research Foundation – Flanders (FWO). The research leading to the results reported in this paper was supported in part by the Research Fund of KU Leuven (GOA/15/003) and by the Interuniversity Attraction Poles programme financed by the Belgian government (IAP/P7/06). The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government – department EWI. This article uses data from the COGITO Study. The principal investigators of the COGITO Study are Ulman Lindenberger, Martin Lövdén, and Florian Schmiedek. Data collection was facilitated by a grant from the Innovation Fund of the President of the Max Planck Society to Ulman Lindenberger. We are grateful to Sylvia Wenmackers for sharing her thoughts on how explanation and prediction are related from a philosophy of science point of view, and to Nicole Geschwind and Marieke Wichers for the use of the MindMaastricht data. The first draft of the manuscript was shared on ResearchGate and presented at the “Networks and time series models” lab meeting in Amsterdam. Kirsten Bulteel and Merijn Mestdagh contributed equally to this work.

Correspondence concerning this article should be addressed to Kirsten Bulteel, Faculty of Psychology and Educational Sciences, KU Leuven, Tiensestraat 102, Box 3713, 3000 Leuven, Belgium. Email: kirsten.bulteel@kuleuven.be

Abstract

In psychology, modeling multivariate dynamical processes within a person is gaining ground. A popular model is the lag-one vector autoregressive or VAR(1) model and its variants, in which each variable is regressed on all variables (including itself) at the previous time point. Many parameters have to be estimated in the VAR(1) model, however. The question thus rises whether the VAR(1) model is not too complex and overfits the data. If the latter is the case, the estimated model will not properly predict new unseen data. As a consequence, it cannot be trusted that the estimated parameters adequately characterize the individual from which the data at hand were sampled. In this paper, we evaluate for current psychological applications whether the VAR(1) model outpredicts simpler models, using cross-validation (CV) techniques to determine the predictive accuracy. As it is unclear whether one should use standard CV techniques (leave-one-out CV or *K*-fold CV) or variants that take time dependence into account (blocked CV, *hv*-block CV, or accumulated prediction errors), we first compare the relative performance of these five CV techniques in a simulation study. The simulation settings mimick the data characteristics of current psychological VAR(1) applications and show that blocked CV has the best performance in general. Subsequently, we use blocked CV to assess to what extent the VAR(1) models predict unseen data for three recent psychological applications. We show that the VAR(1) based models do not outperform the AR(1) based ones for the three presented psychological applications.

Keywords: vector autoregressive modeling; cross-validation; predictive accuracy; within-person dynamics; individual differences

VAR(1) Based Models do not Always Outpredict AR(1) Models in Typical Psychological Applications

Psychologists increasingly study how processes unfold and interact over time at the level of the individual (Hamaker, Ceulemans, Grasman, & Tuerlinckx, 2015). This increase is further fueled by recent calls for a paradigm shift towards reconceiving psychological constructs as networks of interrelated variables (Borsboom & Cramer, 2013; Bringmann et al., 2016). For example, in research on affect, within-person analyses are used to shed light on the relation between positive and negative affective states (e.g., Coifman, Bonanno, & Rafaeli, 2007; Reich, Zautra, & Davis, 2003; Zautra, Reich, Davis, Potter, & Nicolson, 2000; Zautra, Berkhof, & Nicolson, 2002). As another example, fMRI studies map within-person dynamics to capture neuronal interactions (e.g., Roebroeck, Formisano, & Goebel, 2005). In the field of clinical psychology, Sbarra and Allen (2009) investigated the dynamics between sleep disturbances and mood in persons with a major depressive disorder.

On the one hand, this focus on within-person dynamics is supported by theoretical accounts showing that in most cases only intra-individual analyses allow to gain insight in psychological processes. The reason being that cross-sectional results can only be generalized to the level of the individual under very stringent (and often unrealistic) conditions (Molenaar, 2004; Molenaar & Campbell, 2009). On the other hand, intensive longitudinal data are now easy to gather due to technological advances (Bolger & Laurenceau, 2013; Hamaker et al., 2015).

To capture the within-person dynamics, vector autoregressive (VAR) models and its mixed model variants are gaining popularity (e.g., Bos, Hoenders, & de Jonge, 2012; Bringmann et al., 2013; Harrison, Penny, & Friston, 2003; Lodewyckx, Tuerlinckx, Kuppens, Allen, & Sheeber, 2011; Pe et al., 2015; Roebroeck et al., 2005; Rosmalen, Wenting, Roest,

de Jonge, & Bos, 2012; Schmitz & Skinner, 1993; Schuurman, Ferrer, de Boer-Sonnenschein, & Hamaker, 2016; Snippe et al., 2015; van der Krieke et al., 2015; van Gils et al., 2014; Wichers, 2014; Wild et al., 2010; Zheng, Wiebe, Cleveland, Molenaar, & Harris, 2013). Basically, a VAR model consists of a set of equations in which each variable is regressed on all variables (including itself) at previous time points (Brandt & Williams, 2007; Lütkepohl, 2005). A VAR model is thus a multivariate extension of the autoregressive (AR) model, in which a variable is only regressed on a lagged version of itself. In psychology, mostly first order or lag 1 autoregressive models are used, meaning that the regressions go back only one time point (such models are denoted as AR(1) and VAR(1) for AR and VAR models, respectively).

Obviously, VAR models are more complex than AR models because a larger number of parameters needs to be estimated (on the order of J^2 for VAR(1) compared to J for AR(1), where J is the number of variables). Models with more parameters will generally lead to a better fit of the observed data because they allow capturing more particularities of a data set. This statement can be verified by computing the in-sample mean squared error (MSE) between the actual and fitted scores. This is illustrated in Figure 1 using the data presented in Bulteel, Tuerlinckx, Brose, and Ceulemans (2016a). These data comprise eight depression-related symptoms measured on about 100 daily measurement occasions for 28 younger women. We compare the fit of the best fitting person-specific VAR(1) and AR(1) models for the eight symptoms. Two simpler benchmark models (with no within-person dynamics) are also presented: A model with a common (but symptom-specific) mean for all persons and a person-specific (and symptom-specific) mean model. The four models can be ordered from simple to complex on a single continuum (as is done for the abscissa of Figure 1), based on the number of parameters and the nesting structure of the models. The in-sample MSE was computed for the first 90% of the observations of each person only (the reason for this choice

is explained in the next paragraph). The solid line in Figure 1 shows how the in-sample MSE averaged across the persons indeed decreases with increasing model complexity.

The question is however if the best fitting model, the VAR(1) model, is not too complex for the data set at hand. If a model is too complex, it will not only fit the regularities present in the data (i.e., the dynamical relations) but will also fit part of the error, which would be problematic when using the estimated parameters to describe the dynamics of individuals from whom the data were sampled. This problem is called overfitting. In the words of Babyak (2004), " "findings" that appear in an overfitted model don't really exist in the population and hence will not replicate" (p. 411). Overfitting can occur regardless of whether the fitted model is true or wrong at the population level. Indeed, even if we fit the true underlying data generating model to a data set, we may overfit in case the data set contains insufficient measurements relative to the number of parameters to obtain accurate estimates (Babyak, 2004).

Overfitting can be assessed by evaluating how well the model is able to predict new or unseen data (i.e., out-of-sample predictions), that have not been used to estimate the parameters of the model. Returning to our example data set, the dashed line in Figure 1 shows the out-of-sample MSE (averaged over persons) if we use the estimated models to predict the remaining 10% of the observations of each person. Up to an AR(1) model, the graphs of the MSE of these out-of-sample predictions and the one of the in-sample MSE are roughly parallel, implying that the relative performances of the models are nearly equal. The out-of-sample MSE when using a VAR(1) model is substantially larger than the in-sample MSE, however, and clearly worse than that of the AR(1) model. Thus, including some parameters capturing time dynamics seems necessary, as the AR(1) model offers the best out-of-sample predictions. However, the VAR(1) model proves to be too complex for these data: There are too many parameters given the number of measurement occasions. Therefore, error has been

fitted and the VAR(1) coefficients do not adequately estimate the unique direct effects of the variables on each other across time.

The preceding example nicely illustrates the phenomenon of overfitting. Especially highly parametrized models such as the VAR(1) model are prone to overfitting. In addition, the example shows the value of evaluating the predictive accuracy (measured with the out-of-sample MSE) to detect overfitting. In an ideal situation, we fit our models on the observed data, collect new data and evaluate the performance of each model on these new data. The model with the best predictive accuracy is selected. However, in most situations, it is not feasible to wait for future data for testing the model. An attractive way out is offered by cross-validation (CV) techniques.

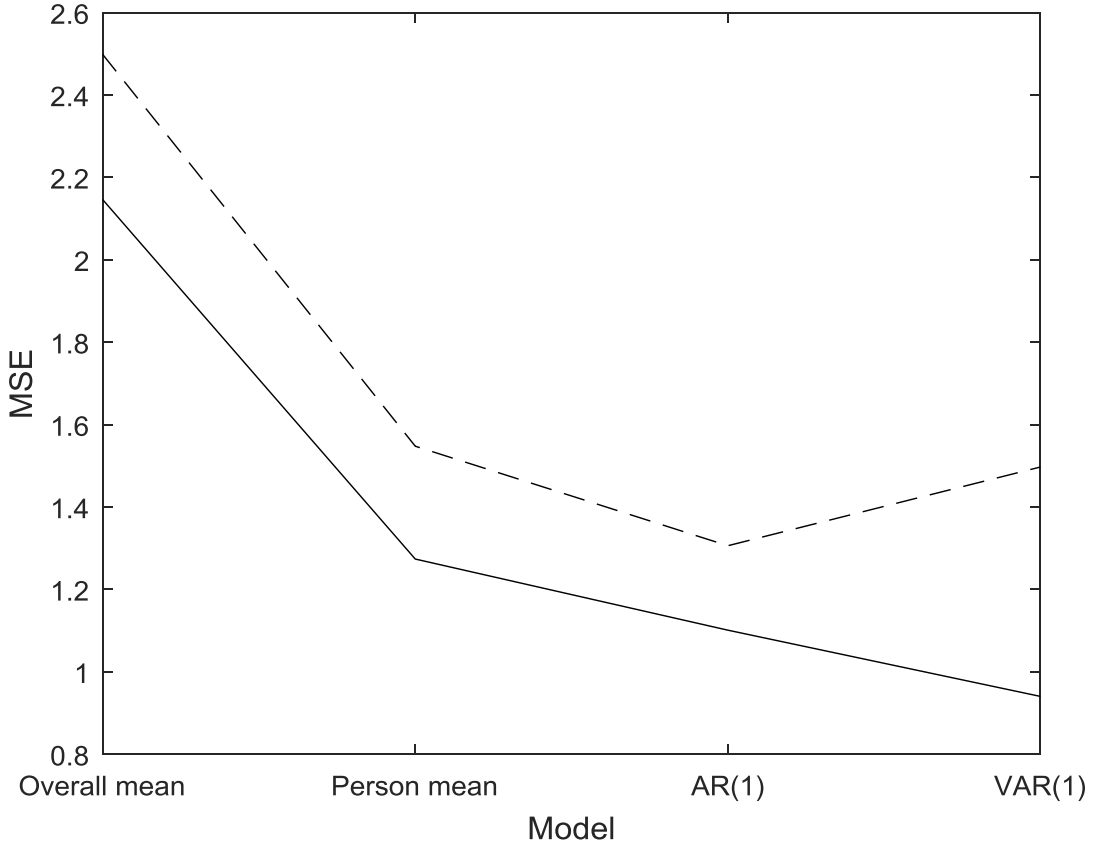

Figure 1. In-sample (solid line) and out-of-sample (dashed line) mean squared errors for four models for the COGITO data in Bulteel et al. (2016a). For more information on the data, see

the text. This Figure is inspired by figures in Hastie, Tibshirani, and Friedman (2009; p. 220), and Pitt and Myung (2002).

CV (Geisser, 1975; Stone, 1974) is a popular and widely applicable method that allows for the estimation of the predictive accuracy based on a single sample. In a typical implementation (called *K*-fold CV), we split the sample randomly into *K* parts of equal size. One part is selected as the test set, and the training set contains the remaining data. The model is fitted to the training set, and then the out-of-sample MSE (i.e., the prediction error for the test set) of this model is computed when predicting the test set. This procedure is repeated by selecting each of the *K*-1 other parts as test set. Finally, the mean of the *K* estimates of the prediction error is calculated (Hastie, Tibshirani, & Friedman, 2009). The model with the lowest average out-of-sample MSE is then preferred. When implementing CV, usually five or ten test sets are chosen. When the size of the test set is one, we speak of leave-one-out CV.

An important assumption underlying CV is that the training and the test set are independent. However, it is difficult to make such an assumption for time series data. Therefore, different modifications to the standard procedure were proposed to remove the dependence in the time series. Blocked CV reduces the dependence by having test sets of consecutive measurements (Snijders, 1988; as cited in Bergmeir & Benítez, 2012). In *h*-block CV (Burman, Chow, & Nolan, 1994) and its asymptotically optimal counterpart *hv*-block CV (Racine, 2000), the training observations that are adjacent in time to the observation(s) of the test set are deleted. Moreover, one may compute accumulated prediction errors (APE; Rissanen, 1986) in which each observation is predicted by building a model on the basis of the previous observations only. Whether one or more of these variants outperforms the others as well as standard CV is not yet clear (Bergmeir & Benítez, 2012).

Given that fitting and interpreting too complex models would weaken rather than strengthen the paradigm shift to within-person dynamics in psychology, it is important to verify the predictive accuracy of the models under consideration, even if one's purposes are purely exploratory. When data are overfitted, parameter estimates do not properly reflect population characteristics rendering it difficult to draw an even tentative statement about the population. The main aim of the paper is to investigate by means of cross-validation to what extent the currently used VAR(1) models generalize to unseen data for three state-of-the-art psychological applications. This aim is inspired by Breiman (2001, p. 204) who states that: "The most obvious way to see how well the model box emulates nature's box is this: … fit the parameters in your model by using the data, then, using the model, predict the data and see how good the prediction is". As it is unclear which CV (related) approach is best used to determine the predictive accuracy, we will also compare the performance of the different approaches in a realistic simulation study in which data sets are generated based on parameter estimates for two of these applications.

Some readers may find it unusual that our study uses prediction to assess which model offers a potential explanation for the relations in the data, as prediction is mostly neglected in psychology. So, let us briefly motivate this choice. We argue that, in addition to goodness of fit (as is measured by in-sample MSE or R^2 type of measures), predictive success is an important criterion when evaluating models for psychological data. As formulated by Shmueli (2010): "A rarer yet important use of data partitioning [e.g., CV] in explanatory modeling is for strengthening model validity, by demonstrating some predictive power" (p. 297). We thus emphasize that although prediction is valuable (Breiman, 2001), in line with the comments of Cox and Efron on Breiman (2001) we do not claim that predictive success is the final objective, as we do not want to sacrifice interpretability by using black box models.

Furthermore, we want to point out that although there is an ongoing debate in the philosophy of science literature concerning the relation between prediction and explanation (i.e., unraveling the underlying mechanism), it is widely accepted that both constitute central goals of any science. Even if explanation is taken to be the main goal of our field, prediction is a partially overlapping goal. Explanations are usually accepted as scientific only if they replicate and thus generate accurate predictions under similar conditions. In a recent review on the changing views on explanation versus prediction, Douglas (2009) argues that they "should not be viewed as competing goals but rather as two goals wherein the achievement of one should facilitate the achievement of the other" (p. 445) and that "it is explanations that produce prediction, which then are successful, that should get our attention" (p. 461-462). Moreover, finding the right trade-off between model simplicity and accuracy is an underlying issue that is relevant for both prediction and explanation. In this paper, we discuss simplicity in relation to avoiding overfitting. Simplicity is also considered an important epistemic virtue in explanations, which should be understandable by us, beings with finite reasoning capacities. In short, we regard our investigation of predictive power partially as supplementary to and partially as supportive for the ongoing search for explanatory models in psychology.

The remainder of the paper is organized as follows. In the following sections, we first define the AR(1) and VAR(1) models (the basic versions and extensions) we are going to investigate in the current paper, and explain the use of CV and APE approaches to obtain a measure of the predictive accuracy of the different models. Next, we describe a simulation study in which we investigate the relative performances of the different CV (related) approaches. Applications to three already published data sets are presented in a fourth section. To conclude, we give a summary of the findings and present directions for future research.

Autoregressive and Vector Autoregressive Models

In this section, we present the models under study. We start with discussing the structure of the data.

Data Structure

The typical data set includes I persons. For each person i, scores on J variables have been observed at T_i measurement occasions (with $i = 1, \ldots, I$). To be able to fit the lagged models through standard regression techniques, each row of the data set contains the J variables of person i at time point t as criterion variables, and the J variables one time point earlier (i.e., the lagged versions of the J variables) as predictor variables. The first measurement of each person is not included in the criterion scores because there is no prediction score one time point earlier available. Similarly, the last measurement of each person is not included in the prediction scores, because no follow-up criterion scores have been gathered. For the time series models that we will use, it is assumed that the intervals between the observations are of equal length (although for some data sets, this is only approximately true).

The Person-Specific AR(1) Model and VAR(1) Model

We can fit an AR(1) or a VAR(1) model to the data of each person separately. In general, the $J \times 1$ vector of observations $\mathbf{y}_{it}$ for person i ($i = 1, \ldots, I$) on time-point t_i ($t_i = 1, \ldots T_i$) is modeled as follows:

$$\mathbf{y}_{it} = \mathbf{c}_i + \mathbf{\Phi}_i \mathbf{y}_{i,t-1} + \mathbf{u}_{it}. \quad (1.1)$$

where $\mathbf{y}_{i,t-1}$ represents the $J \times 1$ vector containing the values of the variables at time point t-1 for person i, $\mathbf{c}_i$ is the $J \times 1$ vector holding the person-specific intercepts, $\mathbf{\Phi}_i$ represents the $J \times J$ matrix of the person-specific regression coefficients, and $\mathbf{u}_{it}$ is a $J \times 1$ vector containing the innovations at time point t. The innovations refer to the part that cannot be predicted based on

the observations at the previous time point. The AR(1) and VAR(1) model are thus a set of regression equations which model each variable as a linear function of the variable scores at the previous time point. The difference between the AR(1) and the VAR(1) model lies in the regression coefficients matrix $\mathbf{\Phi}_i$. For an AR(1) model, we only estimate the diagonal elements (i.e., the effect of a variable on itself), and the off-diagonal elements are set to zero, whereas all coefficients are estimated for a VAR(1) model. Thus an AR(1) model for a set of variables is nested within the VAR(1) model for these data.

The following assumptions are made. First, the innovations follow a multivariate normal distribution with zero means and a variance-covariance matrix $\mathbf{\Sigma}_u$. In case of a VAR(1) model, this implies that the innovations can be correlated at the same time point, but not across time points. Second, stationarity is assumed in that the (joint) distribution of the time series should be time invariant (Lütkepohl, 2005). Therefore, the eigenvalues of $\mathbf{\Phi}$ should have a modulus smaller than 1 (Lütkepohl, 2005). In case of an AR(1) model, this simplifies to the condition that the absolute value of each AR coefficient individually has to be smaller than 1.

To estimate the parameters of an AR(1) or a VAR(1) model, various procedures are available including least squares (LS) estimation methods, Yule-Walker estimation, and maximum likelihood (ML) estimation (see e.g., Hamilton, 1994; Lütkepohl, 2005). LS and ML estimators yield identical estimates (Lütkepohl, 2005). Yule-Walker estimators share the asymptotic properties with LS and ML estimators, but might perform worse in small samples (Lütkepohl, 2005).

Mixed Model Extension of the Person-Specific AR(1) and VAR(1) Models

To improve the estimation of the individual-level parameters of the models presented in the previous section, we can benefit from pooling the information across participants in a mixed model to borrow strength from the other participants:

$$\mathbf{y}_{it} = (\mathbf{c}^{g} + \mathbf{c}_{i}) + (\mathbf{\Phi}^{g} + \mathbf{\Phi}_{i})\mathbf{y}_{i,t-1} + \mathbf{u}_{it}, \quad (1.2)$$

where parameters with superscript *g* give the group-level estimates or the fixed effects. In mixed models, the person-specific deviations $\mathbf{c}_i$ and $\mathbf{\Phi}_i$ are assumed to come from a population distribution (usually a multivariate normal distribution with zero means and a full covariance matrix) and they are called the random effects (Bringmann et al., 2013). Otherwise, the same assumptions as for the person-specific AR(1) and VAR(1) model apply. We will call these models the mixed AR(1) and mixed VAR(1) models. Mixed models are also known as multilevel or hierarchical models.

The random effects distribution implies a shrinkage of the person-specific AR(1) coefficients towards the fixed effects or group-level estimates, which helps to prevent overfitting. Intuitively explained, we have a limited number of measurement occasions, and thus information, for each person. If we want to obtain the best guess for the person-specific AR(1) coefficients, we assume that the information we have about the other persons tells us something about which coefficient values are likely for a particular person. More specifically, the less available information about a person, the closer the best guess for this person will be to the overall AR(1) coefficient for all persons (Hox, 2010). Because the person-specific guesses are partly based on the data of other persons they are less prone to overfitting.

To fit the model, we first estimate the group-level parameters and the covariance matrices for the random effects (i.e., the random effects are integrated out of the model by using their population distribution). To this end, we use a so-called pseudo-likelihood procedure by estimating these parameters for each criterion variable separately (i.e., equation

by equation) because a simultaneous approach is computationally not feasible at this point (Bringmann et al., 2013; Liu, 2017)[1]. As a consequence, not all covariances between the random effects are directly estimated. We infer these parameters based on the residuals. The person-specific deviations (i.e., $\mathbf{c}_i$ and $\mathbf{\Phi}_i$) are found in a second step by using the best linear unbiased predictor.

Lasso VAR(1) Model

Another well-known extension to prevent overfitting is the lasso (least absolute shrinkage and selection operator; Tibshirani, 1996; for applications in time series literature: e.g., Abegaz & Wit, 2013; Hsu, Hung, & Chang, 2008). We start with a stochastic model identical to the one of the VAR models (Equation 1.1), but add a penalty term to the LS loss function or loglikelihood function. This penalty is the sum of the absolute values of the VAR(1) coefficients, weighed with a tuning parameter. While setting the tuning parameter to zero simplifies the lasso VAR(1) model to the regular variant, making the tuning parameter large will effectively set many coefficients equal to zero.

As was the case for the mixed AR(1) models, we use a pseudo-likelihood method that estimates the parameters equation by equation, so for each criterion variable separately. Choosing a value for the tuning parameter is done with CV techniques (see next section). Note that the tuning parameter is determined for each equation individually, and will thus probably differ across equations (Rothman, Levina, & Zhu, 2010, refer to this as a ‘separate lasso’).

Cross-Validation

[1] Bayesian methods can be used for simultaneous estimation, but this may be very time consuming (Schuurman et al., 2016).

Our goal is to determine the predictive accuracy of the models discussed in the previous section. To this end, we use CV to estimate the prediction error for unseen data, by means of the out-of-sample MSE of the different models (from now on, shortened to MSE), computed between the predicted and actual scores of the test data (see introduction). The predictions are computed using the parameters estimated on the training part of the data at hand. The model with the smallest estimated prediction error is selected. Because the prediction error is an estimate and thus subject to uncertainty, Hastie et al. (2009) propose to select the most parsimonious model within the range of one standard error above the prediction error of the best model (this is the one standard error rule).

In this section, we shortly describe different CV (related) approaches. We start with a discussion of the standard techniques. These standard methods are easily applied for iid data, but their application to time series data requires some modification. Thus, we also cover the approaches specific for time series.

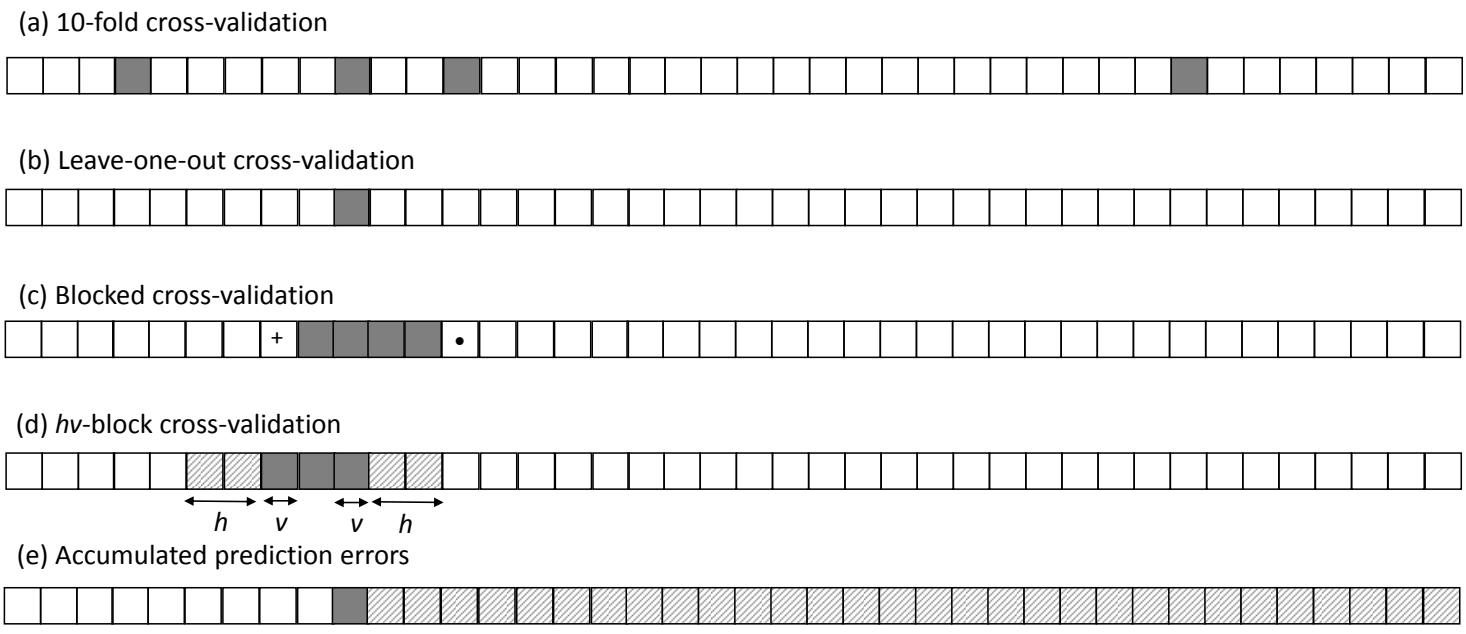

Figure 2. For a data set of 40 observations (represented as squares), one training set, the associated test set, and the discarded elements are shown for different cross-validation (related) approaches. Dark colored squares indicate the elements of the test set. The blank squares are the elements of the training set. Shaded squares indicate measurements that are not considered.

K-fold Cross-Validation

The procedure for *K*-fold CV (Geisser, 1975) is as follows. For each person, the sample is randomly divided in *K* equal parts. One subsample is selected as the test set, and the remaining parts become the training set, as is illustrated in the first panel of Figure 2. The model is fitted to the training set, and the estimated parameters are then used to predict the observations of the test set. For each criterion variable *j* (*j* = 1, …, *J*), the MSE between the predicted and the actual observations of the test set is computed. These steps are repeated using each of the *K*-1 other subsamples as test set. Next, the MSE scores for variable *j* are added over the *K* repetitions (Hastie et al., 2009), and averaged over all available observations in the data set, yielding a weighted average MSE_j with persons with more observations getting a larger weight:

$$\text{MSE}_j = \frac{1}{T}\sum_{i=1}^{I}\sum_{k=1}^{K}\left\|\mathbf{y}_{ijk} - \mathbf{y}_{ijk}\right\|^2, \tag{1.3}$$

where *T* is the total number of time points (i.e., $\sum_{i=1}^{I} T_i$), *K* equals the number of test sets (*K*=5 or 10), $\mathbf{y}_{ijk}$ indicates the *j*th column of the $T_{ik} \times J$ $\mathbf{Y}_{i,k}$ matrix containing the T_{ik} test set observations of the *i*th person in the *k*th fold , and $\mathbf{y}_{ijk}$ denotes the corresponding column of the $T_{ik} \times J$ $\hat{\mathbf{Y}}_{i,k}$ matrix holding the test set scores predicted based on the training set, and with

$\|\cdot\|^2$ being the squared norm. The standard error of MSE_j is computed as the standard deviation of all $\left\|\mathbf{y}_{ijk} - \mathbf{y}_{ijk}\right\|^2$ scores, divided by the square root of the number of observations T.

Leave-One-Out Cross-Validation

Another classical approach, although computationally more demanding, is leave-one-out CV (Geisser, 1975; Stone, 1974). It can be considered equivalent to a K-fold CV, with $K = T_i$. Hence, the test set consists of the observations on a single time point (as can be seen in the second panel of Figure 2).

Blocked Cross-Validation

Blocked CV is a variant of K-fold CV, in which each of the K blocks now only contains consecutive measurements (Snijders, 1988; as cited in Bergmeir & Benítez, 2012). An illustration can be found in the third panel of Figure 2. To comply with the assumption of equidistant lag intervals, we do not use the observations right before the deleted block (the square with a plus sign in panel c of Figure 2) to predict the first observations after the deleted block (the square indicated with a circle).

***hv*-Block Cross-Validation**

Racine (2000) proposed and studied the *hv*-block CV as a CV method for dependent data. This method is consistent in that the probability of selecting the optimal model in a MSE sense converges to one as the number of observations goes to infinity. As is the case for leave-one-out CV, we select each observation in turn as the test set. In addition, as can be seen in the fourth panel of Figure 2, v observations before and after the selected one are added to the test set to ensure consistency (for more details: Racine, 2000; Burman et al., 1994), and h observations on either side of the test set are discarded from the training set to remove the sequential dependence. If v and h equal zero, the approach reduces to leave-one-out CV.

Similar to blocked CV, the test sets contain a number of consecutive measurements. However, in *hv*-block CV some observations on either side of the test set are removed, and each observation is included in multiple test sets. As was the case for the previous method, we do not use the observations right before the deleted block to predict the first observations after the deleted block to respect the assumption of equidistant lag intervals in the training set.

With regard to specifying the *h*- and *v*-hyperparameter, recommendations can be found in Racine (2000). However, the choice will also depend on the data set at hand to ensure that the training set contains sufficient observations to estimate the most complex of the fitted models.

Accumulated Prediction Errors

An approach related to the CV techniques was proposed by Rissanen (1986). It takes the sequential ordering of the time points into account by using only observations earlier in time to make predictions. APE is analogue to leave-one-out CV but removes all observations from the training set that occurred after the test set observation (as can be seen in the final panel of Figure 2). To be able to fit the most complex model, it is recommended to define a minimal length k_{APE} of the training set, implying that the first k_{APE} time points are never used as test observation.

Simulation Study

Research Questions

Before we can evaluate the predictive accuracy of the different time series models discussed above in real data, we first study the performance of the different CV (related) approaches. Therefore, we examine both their ability to estimate the prediction error, and their

ability to select the best model in a predictive sense. Second, we link the true prediction error to parameter accuracy.

In order to interpret the predictive performance of the time series models discussed above (person-specific AR(1) and VAR(1), mixed AR(1) and VAR(1), and lasso VAR(1)), we add a few benchmark models. In particular, the overall mean model, the person mean model, and a mixed model having only random intercepts and no slopes (referred to as the mean mixed model; note that this model is sometimes called an empty model in the mixed modeling literature) are added because they do not allow for time dependence. By comparing the predictive accuracy of models with and without AR effects, we want to assess whether adding time dependency improves the quality of the predictions. For the overall mean model, the predicted scores for a person equal the mean of the variables across persons in the training set. In case of a person mean model, the mean values of each particular person are used to predict the corresponding scores. The difference in predictive accuracy for the overall versus the person mean model allows to examine whether incorporating individual differences is useful in a predictive sense.

Design and Procedure

To generate realistic time series data, we used parameter values obtained from analyzing real data sets. In particular, we reanalyzed the data set in Bringmann and colleagues (2013; for more details see Geschwind, Peeters, Drukker, van Os, & Wichers, 2011) and a subsample of the COGITO dataset (as reported in Bulteel et al., 2016a; for more details see Schmiedek, Lövdén, & Lindenberger, 2010) which we also discuss in the Applications section below, and hereafter refer to as the MindMaastricht data and the COGITO data, respectively. The MindMaastricht data originates from a typical experience sampling method (ESM) study in which participants answered a questionnaire about six variables several times a day for a

number of days. In the COGITO study participants filled out a daily questionnaire on eight symptoms for about 100 days. We also include the latter data set in the simulation study because the time dependence is expected to be considerably smaller as participants are measured only once per day.

We start with estimating the parameters of the following models for these observed data: (a) An overall mean model, (b) a person mean model, (c) a mean mixed model, (d) a person-specific AR(1) model, (e) a person-specific VAR(1) model, (f) a mixed AR(1) model, (g) a mixed VAR(1) model, and (h) a lasso VAR(1) model. Blocked CV is used to determine the tuning parameters for the lasso VAR(1) model. Model parameters can be estimated with standard MATLAB functions.

Using each of the 16 sets of estimated parameters (2 data sets × 8 generating models), we simulated 100 data sets, resulting in 1600 unique data sets[2]. For simulations based on the MindMaastricht data, we generated a time series of 41 observations (i.e., the average number of time points per person in the original data set) for each of the 52 persons. For simulations based on the COGITO data, the time series consist of 70 observations for each of the 28 persons; note that we used 70 rather than 100 time points because Bulteel et al. (2016a) deleted on average 30 time points per person to ensure equal time intervals between the time points. To simulate the time series of a particular person, the initial values were the person-specific average scores in the original data set. A burn-in period of 1,000 observations was used to remove the influence of the starting values and to obtain stationary time series[3].

To check if the conclusions apply to longer time series as well, we also included an additional condition for the MindMaastricht data in which each person has a time series of

[2] The lasso VAR(1) model is not a data generating model as such; we fit the lasso VAR(1) model to the observed data and then use the VAR(1) model with the estimated parameters to simulate new data.
[3] An alternative for the use of a burn-in period is to sample time points from the stationary distribution for each model.

500 observations. We simulated 10 data sets for this condition, as longer time series are computationally more demanding especially in an extensive simulation study (e.g. for a leave-one-out CV we need to fit each model 500 times). For the MindMaastricht data, we thus have 110 data sets (100 short and 10 long) for each of the 8 generating models.

Next, we applied the five CV (related) approaches to determine the MSE (averaged across persons) for the eight estimated models per simulated data set (i.e., resulting in 40 overall MSE values per simulated data set, as can be seen in Figure 3): (a) 10-fold CV, (b) leave-one-out CV, (c) blocked CV, (d) *hv*-block CV, and (e) APE. Regarding the use of the *hv*-block approach, we choose *h* and *v* such that the resulting test set had a size comparable to that of 10-fold CV and blocked CV. More specifically, $h = 1$ and $v = 2$ for the MindMaastricht data with $T = 41$; $h = 10$ and $v = 57$ for the MindMaastricht data with $T = 500$; and $h = 2$ and $v = 5$ for the COGITO data[4]. The minimum training set length k_{APE} for the APE is set to 10, and to 100 for the $T = 500$ condition.

The performance of the CV approaches was then examined based on two indicators. The first indicator is the accuracy of the overall MSE measure, which is obtained by averaging all variable-specific MSE_j values, as an estimate for the true prediction error. This true prediction error is defined as the average prediction error that results from, first, fitting a model on the complete observed data set at hand (and hence not carving it up in training and test sets), and, second, using the estimated parameters to predict randomly sampled unseen data of the same individuals on the basis of their (also unseen) scores one time point earlier (assuming that the data generating mechanism remains the same). Because the true prediction error is unknown, we approximate it by creating a very long time series (i.e., $T_i = 10{,}000$ for *i*

[4] We also set *h* and *v* as close as possible to the values recommended by Racine (2000) while retaining at least 30 observations in the training set. The results were however substantially worse than the ones for the reported *hv*-block CV, and are therefore not discussed here.

= 1, …, *I*) for each of the 16 data generating settings (2 data sets x 8 generating models) in our simulation, that take the role of the randomly sampled unseen data mentioned above, for each of the 100 (COGITO) or 110 (MindMaastricht) data sets and all 8 estimated models (columns of the table in the third step of Figure 3), yielding 800 (COGITO) or 880 (MindMaastricht) true prediction errors. We then want to investigate which CV technique (i.e., the rows in the table of step 3) approximates this true value best by computing the mean squared difference between the MSE values and the true prediction error. We will refer to this difference as the error of CV.

Our second performance indicator is the percentage of data sets for which the best model in a predictive sense was chosen, that is, the model for which the true prediction error is minimal. Importantly, the data generating or true model is not by default the best predictive model. One reason for this discrepancy is that the data characteristics (e.g., a relatively low number of time points per person) might impede proper estimation of a true but more complex data generating model, leading to poor predictions. Of course, if one were to have an infinite amount of noise-free data, the true model and best predictive model would coincide. Thus, summarizing, when CV shows that a candidate model has a low predictive accuracy compared to other models, this can mean two things: A first option is that the candidate model simply is not the correct data generating model, and therefore generating bad predictions. A second option is that the model is correct, but the data are not informative enough or too noisy to sufficiently accurately estimate its parameters, again leading to bad predictions.

Lastly, we also investigated how the true prediction error is related to parameter accuracy. To this end, we first studied the problem theoretically in the simple context of an AR(1) model. It can be shown (derivation is given in the Supplemental Materials) that for a simple AR(1) model without intercept, the relation between the prediction MSE (averaged over new unseen data) and the parameter accuracy is given by:

$$E(\text{MSE}) = \left(\beta - \tilde{\beta}\right)^2 \frac{\sigma_u^2}{1-\beta^2} + \sigma_u^2,$$

where β is the true AR parameter, $\tilde{\beta}$ is its estimate based on a training sample, and σ_u^2 is the variance of the innovations. A first observation is that (given the AR(1) model is correct) the parameter accuracy $\left(\beta - \tilde{\beta}\right)^2$ will go to zero when the number of observations T_i grows large. Consequently, $E(\text{MSE})$ will converge to its irreducible minimum σ_u^2 , because you cannot predict better than the innovation variance. Let us now define the surplus mean square error (the amount of prediction error on top of the irreducible minimum) $\text{MSE}^* = \text{MSE} - \sigma_u^2$. It holds that:

$$\log E(\text{MSE}^*) = \log\left(\beta - \tilde{\beta}\right)^2 + \log\frac{\sigma_u^2}{1-\beta^2},$$

which shows that for a simple AR(1) model, the log-surplus error and the log parameter accuracy are a simple shift of one another, where the size of the shift depends on the innovation variance and the AR parameter. Thus, if we would plot the log-surplus error and the log parameter accuracy against the number of observations, both curves decrease parallel.

In a next step, we studied the same relation in a more realistic context. We estimated the mixed VAR(1) model for the COGITO data set. We then used the obtained model parameters to simulate 100 training data sets of nine different lengths: $T_i = \{2^2, 2^3, \ldots, 2^9 = 512\}$, for *i* = 1, …, *I*. We also generated one very long test data set with $T_i = 10000$. For each of the 900 training data sets we then estimated a mixed AR(1) model and a mixed VAR(1) model. As we simulated each data set, we know the true model parameters, that is the coefficients $\Phi^g + \Phi_i$ of Equation 1.2. Therefore, we can now also compute the parameter accuracies (MSE of the parameter estimates excluding the intercepts), and compare this to MSE^* (the MSE of the predictions for the long test data set with the innovation variance subtracted).

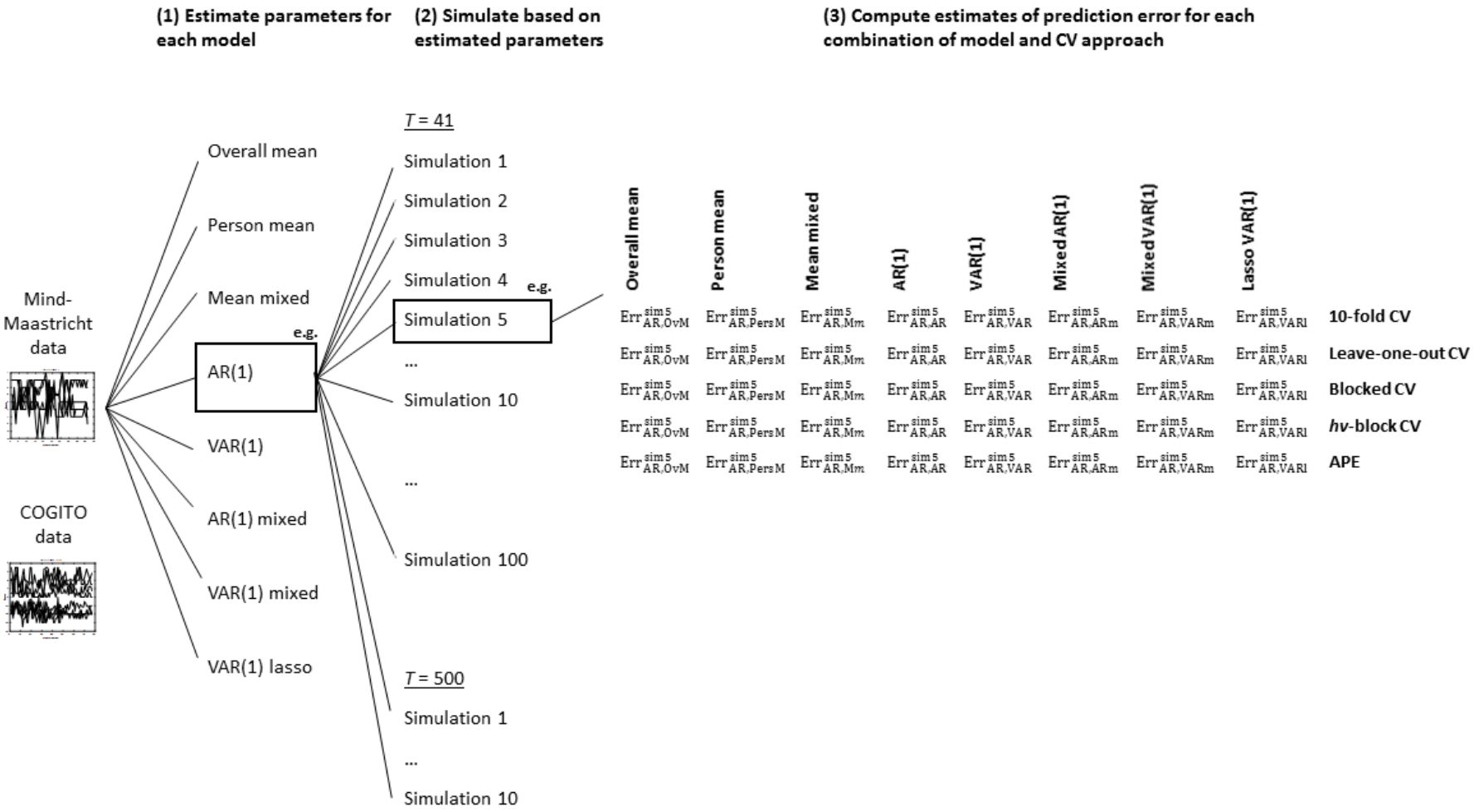

Figure 3. Visualization of the simulation procedure. At the left side, the two observed data sets (MaastrichtMind or COGITO) are shown on which the simulation study is based. Step 1 refers then to the fitting of the eight simulation models to such an observed data set and extracting the parameter estimates. In a second step, 100 data sets are simulated based on the previously estimated parameters (i.e., simulations 1 to 100) for all

eight simulation models; for the MindMaastricht data 10 longer time series are simulated as well. In a third step, eight estimation models are fitted (this time to the simulated data, e.g., simulation 5) and five CV methods are used to compare the predictive accuracy of the fitted models. Several prediction errors are estimated. For example, $Err^{sim5}_{AR,OvM}$ is the estimated prediction error when analyzing the fifth simulated data set from an AR(1) model with an overall mean model.

Results

First, we evaluate how well the different CV approaches estimate the true prediction error. We first inspect the 100 data sets with a length similar to the original ones. Figure 4 shows, for the simulated data based on the MindMaastricht data, the error of CV for each combination of CV approach and the model used to estimate the parameters averaged across the 100 generated data sets, separately for each model used to simulate the data. The APE estimate of the prediction error is considerably worse than for the other CV approaches and this is most dramatic for the person-specific VAR(1) model (regardless of the model used to generate the data). Also the *hv*-block CV method misestimates the true prediction error for the person-specific VAR(1) model (albeit to a lesser extent than the APE). The main reason is that given that the time series are rather short per person, discarding a substantial number of data points (as is done in APE and *hv*-block CV), renders it impossible to fit the highly parametrized VAR(1) model properly.

When the simulated data contain no time dependence, the leave-one-out CV performs best (panels a, b, and c of Figure 4). When time dependence is present, the standard CV approaches perform worse than the modified approaches that take time dependence into account and especially have difficulties to estimate the prediction error for the person mean model. Out of the modified approaches, blocked CV is the best option in case of time dependency. As blocked CV performs only slightly worse than leave-one-out CV when no time dependency is present, it is the best performing method in general because it has the lowest overall error of CV (as can be seen in Table 1). Similar conclusions hold for the COGITO data. The relative results for the condition with $T = 500$ for the MindMaastricht data are in line with the $T = 41$ condition, but the absolute error of CV values is substantially lower.

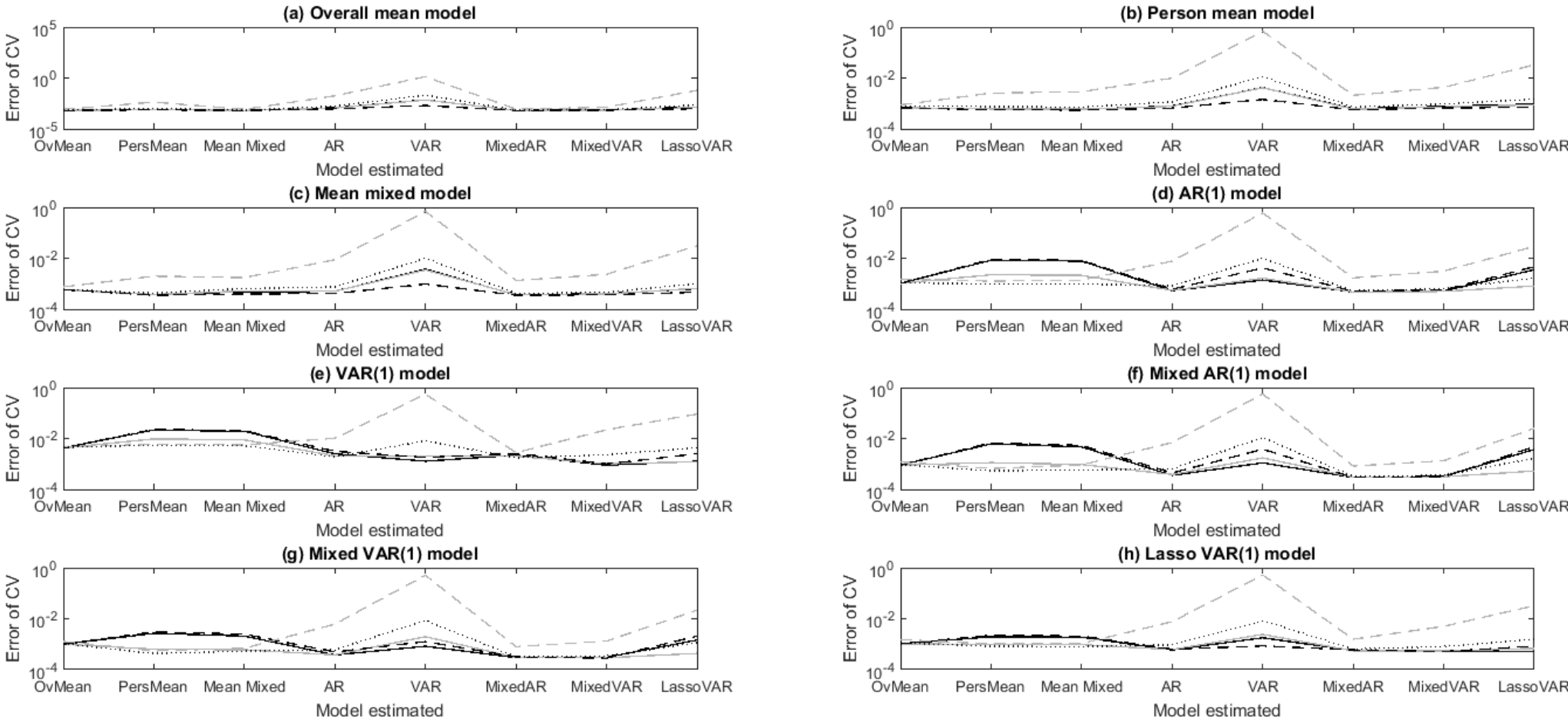

Figure 4. The error of cross-validation for the MindMaastricht data (T = 41 condition). The different panels refer to the data generating models. Note that the *y*-axis is in log scale to make the differences between the methods clearer. The results for 10-fold cross-validation are indicated with a solid black line, for leave-one-out cross-validation with a dashed black line, for blocked cross-validation with a solid gray line, for *hv*-block cross-validation with a dotted black line, and for accumulated prediction errors with a dashed gray line.

Table 1

The estimation error of various CV methods of the true prediction error averaged across data generating models, fitted models, and the 100 simulated data sets (or 10 simulated data sets for T=500).

	T	**10-fold CV**	**Leave-one-out CV**	**Blocked CV**	***hv*-block CV**	**APE**
MindMaastricht data	41	2.1811×10^{-3}	2.2226×10^{-3}	1.4649×10^{-3}	2.5679×10^{-3}	87.317×10^{-3}
	500	6.2486×10^{-5}	6.1107×10^{-5}	5.4341×10^{-5}	7.7292×10^{-5}	19.533×10^{-5}
COGITO data	70	10.930×10^{-4}	10.768×10^{-4}	8.9023×10^{-4}	16.874×10^{-4}	22309×10^{-4}

Table 2 shows the percentage of simulated data sets for which the best model in the predictive sense was selected by the CV (related) approaches (i.e., the model with the lowest true prediction error), separately for both original data sets. Again, the differences between the CV approaches are rather small. For the MindMaastricht data, blocked CV performs best: In 89% of the cases, the model with the lowest true prediction error was also identified by the blocked CV. The remaining CV approaches perform only slightly worse. The APE performs substantially worse (77%). In the $T = 500$ condition, leave-one-out CV has the best performance (81%), but is closely followed by blocked CV (80%). Note that the percentages are lower on average for this condition, because the predictive accuracy of the person-specific and mixed variant of the same model are very similar for the $T = 500$ condition. The reason is that borrowing strength from the other persons in mixed models is not advantageous anymore,

in case we have sufficient data to accurately estimate the model for each person separately. Therefore, the performances of the person-specific and mixed models will be almost identical, and a decision between both models is no longer required. Indeed, if we do not distinguish between the person-specific and mixed variants of the same model, blocked CV and leave-one-out CV select the best model in the predictive sense in 92.5% of the cases. For the COGITO data, blocked CV also performs best (83%), but the performance of leave-one-out and *hv*-block CV is nearly as good (i.e., 81% and 80% respectively).

One may wonder how often the best model in the predictive sense equals the data generating model. For the data sets with a number of time points per person similar to the one of the original data sets, the best model in the predictive sense was always the mixed variant of this model in case the data were simulated with a person-specific AR(1) or VAR(1) model. This result probably follows from the rather limited number of time points, and the strength of regularization in the mixed models. Also for the other models, the mixed variants were often the best model in the predictive sense. By inspecting the results for the condition with the longer time series ($T = 500$), we can verify whether this is indeed due to the limited number of time points. In this condition, person-specific models can be the best model in the predictive sense. In particular, the mixed AR(1) model was the best predictive model in 60% of the data sets simulated with the person-specific AR(1) model, and the mixed VAR(1) model only for 10% of the data sets simulated with the person-specific VAR model.

Table 2

Percentage of data sets for which the best model in the predictive sense was selected

	T	10-fold CV	Leave-one-out CV	Blocked CV	*hv*-block CV	APE
MindMaastricht data	41	86%	88%	89%	88%	77%
	500	71%	81%	80%	78%	74%
COGITO data	70	74%	81%	83%	80%	66%

In Figure 5, we show that, for the person-specific AR(1) and VAR(1) models, and the mixed AR(1) and VAR(1) models, the accuracy of the parameters is indeed strongly related to the MSE of the predictions based on simulations from the mixed VAR(1) model. As was the case for the theoretical analysis of the simple AR(1) model, also for the person-specific VAR(1) model and the mixed models, the log-accuracy of the parameters (MSE of the parameters excluding the intercept, averaged across all subjects, variables, and training sets[5]) runs parallel with the log-MSE* of the predictions across the different lengths. For these models, it thus holds that $\text{MSE}^* \approx C \cdot \text{Acc}$ with Acc referring to the parameter accuracy and *C* being a constant. Furthermore, we see that, as expected, the prediction MSE and the parameter accuracy become better if the number of observations *T* of the time series increases. Comparing the person-specific models, it can be seen that for low *T*, the person-specific AR(1) model performs better and the person-specific VAR(1) model overfits; for larger values of *T*, the person-specific VAR(1) model performs better. Similar findings can be found for the comparison between the mixed AR(1) model and mixed VAR(1) model. However, fewer observations are needed for the mixed VAR(1) model to outperform the mixed AR(1) model, compared to the person-specific models. Importantly, both for the person-specific and for the mixed model case, the number of observations *T* that is needed for VAR(1) to become

[5] The off-diagonal elements of $\widetilde{\mathbf{\Phi}}$, the estimated variables for the mixed AR(1) model, are 0.

better than AR(1) is the same for the prediction MSE as well as for the parameter accuracy. This means that, if cross-validation results indicate that one model leads to better predictions, its parameter values are probably also more trustworthy. In addition, Figure 5 shows that the person-specific and mixed variants perform roughly equal at large *T*. The shrinkage of the mixed models is indeed no longer an advantage in case a large number of observations are available. Note that the *T* values at which one model becomes better than another model only hold for this simulation study, and do not generalize to empirical data and thus also not to the applications presented below.

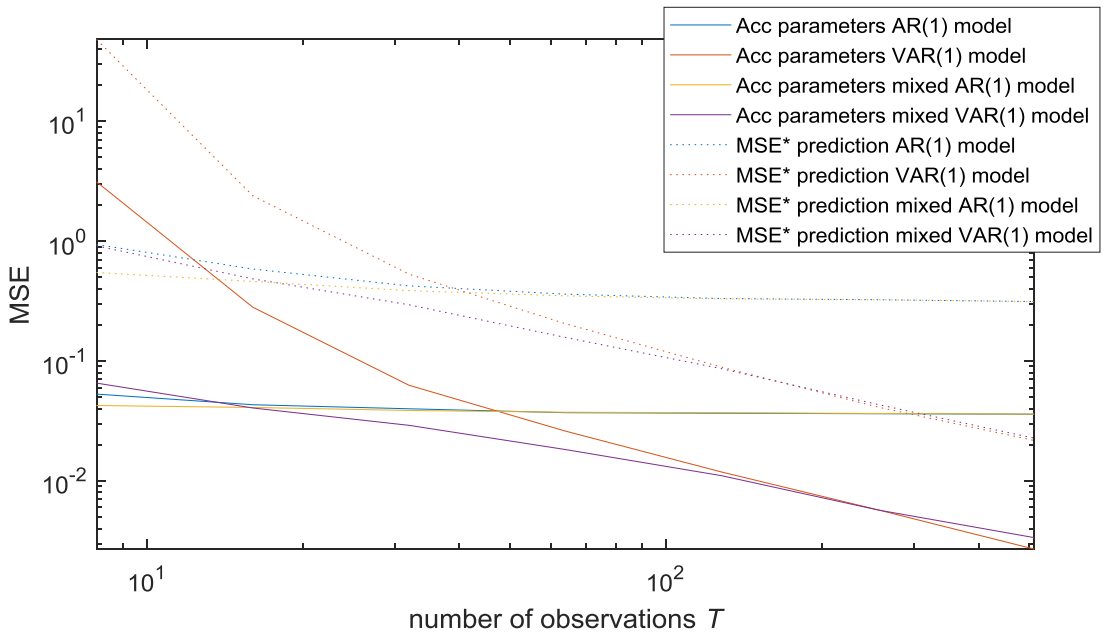

Figure 5. The accuracy of the parameters and the MSE^* of the predictions for the person-specific AR(1) and VAR(1) models, and the mixed AR(1) and VAR(1) models for different numbers of observations *T*. The data sets are simulated based on the estimated mixed VAR(1) model for the COGITO data set. These results are based on 100 simulations for each value of *T*.

Conclusion

Given its good performance in estimating the prediction error as well as in selecting the best model in a predictive sense, we recommend to use the blocked CV approach, thereby following the recommendation of Bergmeir and Benítez (2012). Advantages of the blocked approach are that it requires a simple modification of the standard 10-fold CV approach (i.e., the test set contains consecutive time points), that it is computationally fast (similar to 10-fold CV), and that no tuning parameters (as for *hv*-block CV) need to be set.

Applications

In this section, we evaluate the predictive accuracy of the eight different models under study (i.e., the overall mean model, the person mean model, the mean mixed model, the person-specific AR(1) model, the person-specific VAR(1) model, the mixed AR(1) model, the mixed VAR(1) model, and the lasso VAR(1) model), when applied to three state-of-the-art psychological applications focusing on within-person dynamics. Two of these data sets (the MindMaastricht data and the COGITO data) have already been discussed above. The third data set is analyzed by Bringmann et al. (2016), and is called below the Assessment data. Based on the results of the simulation study, we use blocked CV to estimate the predictive accuracy.

All three data sets were analyzed with VAR(1) models before (Bringmann et al., 2013; Bringmann et al., 2016; Bulteel et al., 2016a). We approximated the original data analysis procedure as closely as possible. However, one additional participant selection criterion was added for the present analysis: Participants should have observations on at least 30 measurement occasions. The cutoff of 30 occasions was introduced to retain sufficient participants on the one hand and to have enough time points per person to fit the models on the other hand.

MindMaastricht Data

We will reanalyze a selection of the MindMaastricht data (Geschwind et al., 2011), reported on by Bringmann and colleagues (2013). We refer to these papers for more detailed information. In short, we analyzed 88 persons with residual depressive symptoms who participated in an ESM study during two periods of six days, with two to three months in between. As we are not interested in the therapy-effect in this paper, the current analyses are limited to the pre-treatment period. Each day of the period was divided in 90-minutes block between 7:30am and 10:30pm. The first measurement of each day was discarded from the criterion scores to avoid over-night prediction. Study participants were notified at a random time in each block, and requested to fill out a questionnaire. Amongst other items, the following six items were answered: 'I feel cheerful', 'I feel relaxed', 'I feel fearful', 'I feel sad', 'worry', and 'pleasantness of the event'. The latter item was an indication of the pleasantness of the most important event that happened between the previous and the current beep. Bringmann et al. (2013) used a mixed VAR(1) model to estimate parameters both at population and at individual level. Subsequently, the estimated parameters were the input of a network analysis.

All estimated models correspond to stationary processes according to eigenvalue analysis, i.e., the modulus of the eigenvalues of the estimated individual VAR(1) processes are smaller than one for all individuals in all models. To give an indication of the size of the obtained (V)AR coefficients, the average absolute value of the parameter estimates, and its 95% interval, can be found in Figure 6 for the different models, separately for the diagonal and off-diagonal elements. The diagonal elements of the AR(1) are around 0.3, and the off-diagonal elements of the AR(1) model are zero by definition. The diagonal elements of the VAR(1) model are slightly smaller than those of the AR(1) model, and the size of the off-diagonal elements is similar to the diagonal ones. The absolute values of the mixed AR(1)

coefficients are similar to the ones of the AR(1) model but the variability is smaller. Comparing the estimates of the mixed VAR(1) to the ones of the person-specific VAR(1) model, mainly the absolute values of the off-diagonal elements, and the variability of both the diagonal and off-diagonal elements is shrunken. For the lasso VAR(1) model, the diagonal elements are substantially shrunken as well, but the variability is larger than is the case with the mixed VAR(1) model.

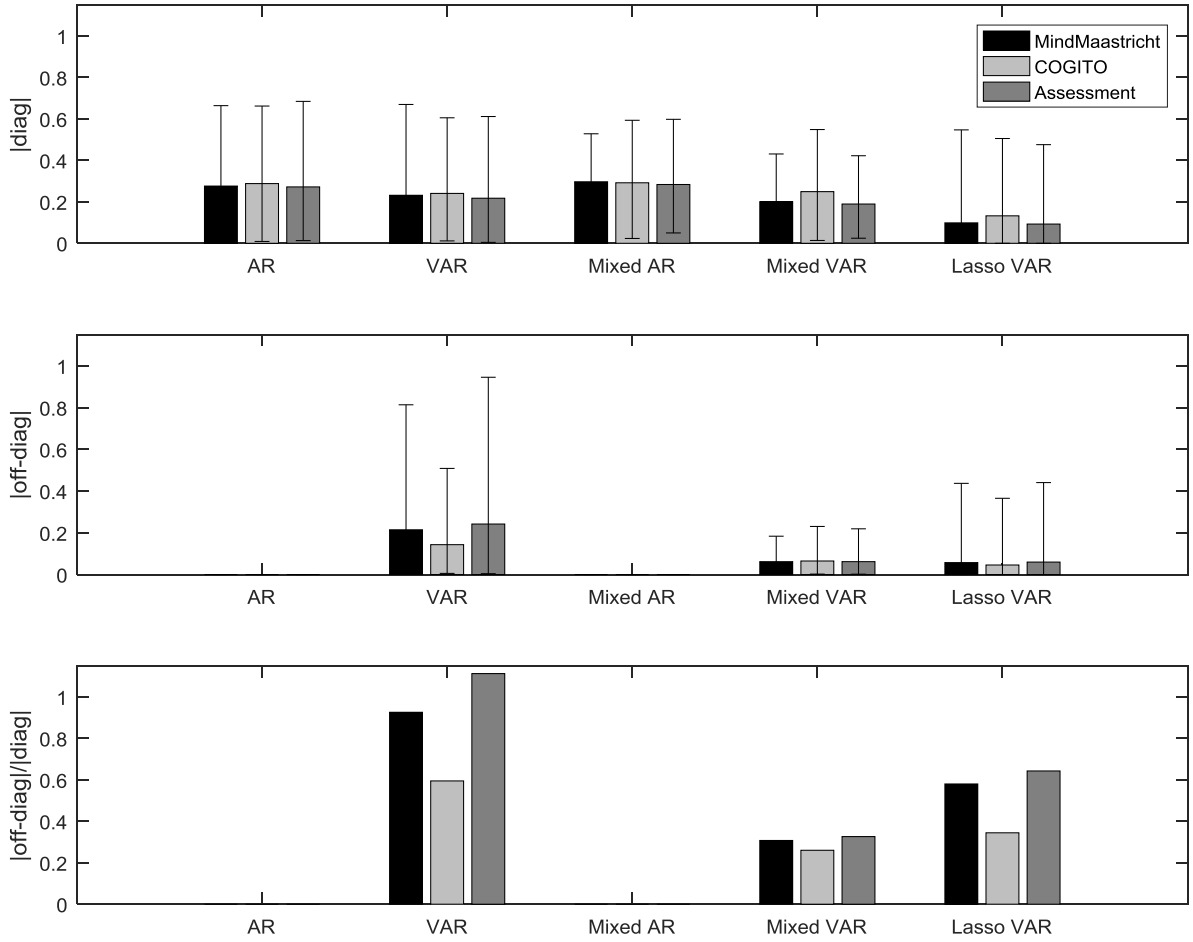

Figure 6. The average absolute values of the off-diagonal (panel b) and diagonal (panel a) parameter estimates and their ratio (panel c) for the three data sets under study. To give an

indication of the variability, the vertical lines in panels a and b show the difference between the 2.5th and 97.5th percentiles.

Table 3 shows the prediction error estimates for all models. The first row of Table 3 displays the prediction error (as estimated by the CV methods) averaged across all six variables, and the remaining rows present the estimates for each variable separately. The largest improvement of the prediction error is obtained when accounting for individual differences, as is clear from the difference between the error estimates for the overall mean model and the person mean model. Including information on the time dependence further improves the predictions except for the most complex person-specific VAR(1) model, which has the second largest prediction error. The predictions based on this person-specific VAR(1) model are worse than the predictions based on a person mean model (or mean mixed model). The overfitting of the person-specific VAR(1) model might also explain the unusual finding that the off-diagonal elements have about the same size as the diagonal elements (see Figure 6). The reason is that the predictive accuracy and parameter accuracy are strongly related, as was shown in Figure 5. The mixed AR(1) and VAR(1) models succeed best in preventing overfitting. As a parsimonious model is to be preferred given the one standard error rule, we select the mixed AR(1) model as the best model for this data set. To further examine our results, we plotted the distributions of the person-specific VAR(1) coefficients in Figure 7. This Figure shows that the distributions of the person-specific parameters are unimodal and symmetric which might explain the superior performance of the mixed model. The third best option is the person-specific AR(1). The lasso VAR(1) models slightly outperform the models without time dependence.

Table 3

The prediction error estimates when applying blocked CV to the MindMaastricht data. The first row is the average MSE across variables and the other rows the MSE per variable. The standard error of the estimates can be found in parentheses

	Overall mean	**Person mean**	**Mean mixed**	**AR**	**VAR**	**Mixed AR**	**Mixed VAR**	**Lasso VAR**
Average	2.46 (0.02)	1.90 (0.02)	1.90 (0.02)	1.77 (0.02)	2.20 (0.03)	1.71 (0.02)	1.71 (0.02)	1.86 (0.02)
Cheerful	2.32 (0.05)	1.75 (0.04)	1.74 (0.04)	1.56 (0.04)	1.99 (0.07)	1.51 (0.04)	1.52 (0.04)	1.63 (0.04)
Pleasantness of the event	2.93 (0.07)	2.80 (0.07)	2.77 (0.07)	2.81 (0.07)	3.54 (0.1)	2.72 (0.07)	2.71 (0.07)	2.91 (0.08)
Worry	3.33 (0.06)	2.23 (0.06)	2.22 (0.05)	1.97 (0.05)	2.47 (0.08)	1.93 (0.05)	1.93 (0.05)	2.13 (0.06)
Fearful	1.46 (0.06)	1.06 (0.04)	1.06 (0.04)	0.98 (0.04)	1.17 (0.05)	0.95 (0.04)	0.95 (0.04)	1.05 (0.05)
Sad	2.47 (0.06)	1.73 (0.05)	1.72 (0.05)	1.55 (0.05)	1.87 (0.06)	1.48 (0.05)	1.48 (0.05)	1.66 (0.05)
Relaxed	2.25 (0.05)	1.87 (0.04)	1.86 (0.04)	1.75 (0.04)	2.19 (0.09)	1.68 (0.04)	1.67 (0.04)	1.78 (0.04)

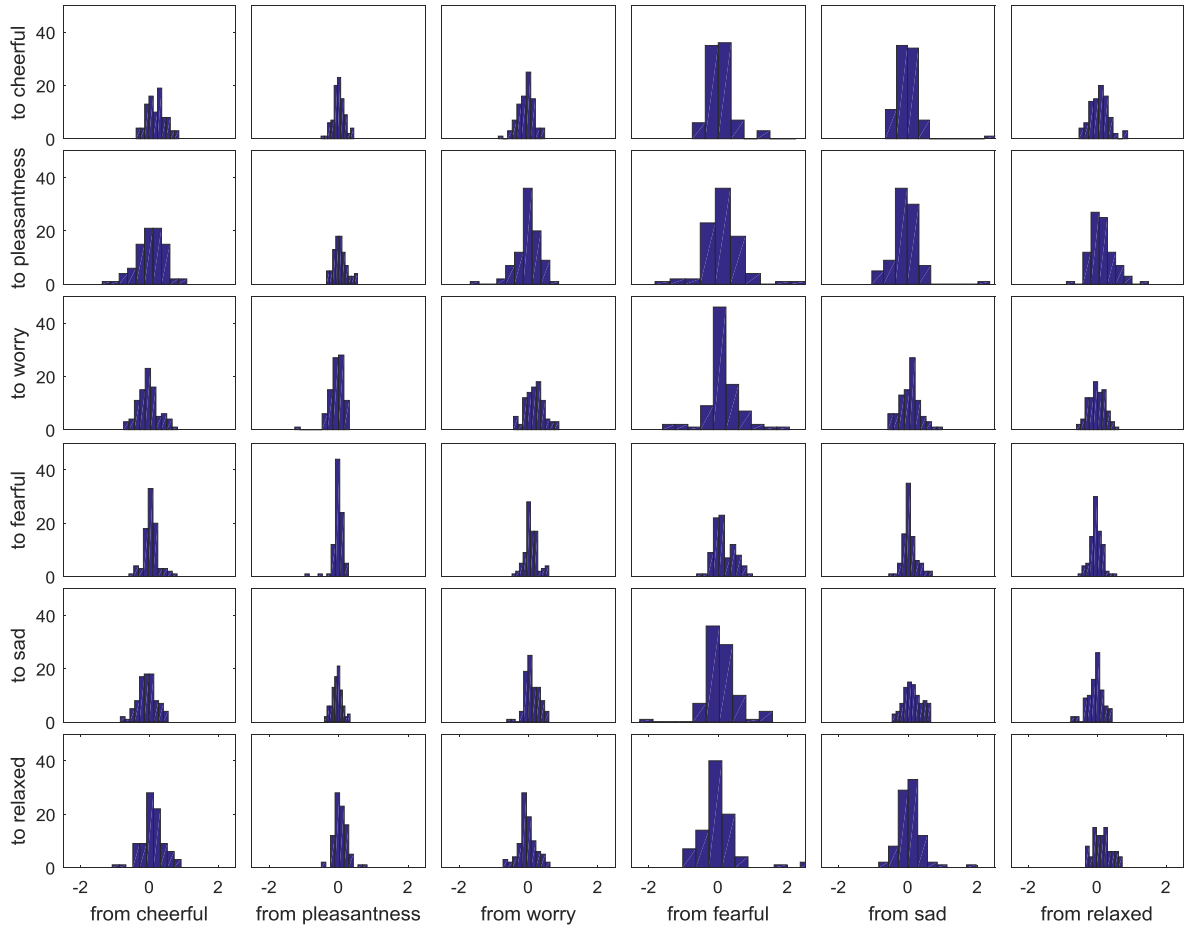

Figure 7. The distribution of the person-specific VAR(1) coefficients for the MindMaastricht data set.

COGITO Data

The second data set is a subsample of the COGITO data (Schmiedek et al., 2010), analyzed in Bulteel et al. (2016a). The subset contains about 70 measurements for 28 younger women. Eight depression-related symptoms are included: Rumination (measured with an 8-point scale), feeling guilty (8-point scale), feeling unhappy (8-point scale), feeling downhearted (8-point scale), loss of activation (8-point scale), loss of interest (8-point scale), sleep quality (8-point scale), and loss of energy (4-point scale). Bulteel et al. (2016) used a

clusterwise VAR(1) model to shed light on the within-person symptom dynamics[6]. More information on the COGITO study can be found in Schmiedek et al. (2010), and on the data selection in Bulteel et al. (2016a). A VAR(1) analysis of one of the participants can be found in Bulteel et al. (2016b).

Figure 6 again shows the absolute values of the parameter estimates for this data set. All estimated models correspond to stationary processes according to eigenvalue analysis. For the regular AR(1) and VAR(1) model, the absolute values of the estimates lie between 0.2 and 0.3. The shrinkage resulting from using the mixed models and the lasso VAR(1) model follows a pattern similar to the one for the MindMaastricht data. Inspecting the prediction error results in Table 4, the largest decrease in prediction error occurs when applying a person-specific model instead of an overall mean model. The models with the lowest prediction error estimates are again the mixed AR(1) and VAR(1) models, with the AR(1) mixed model being preferred according to the one standard error rule. The lasso VAR(1) model has a similar but substantially worse performance than the mixed AR(1) or VAR(1) models. The person-specific VAR(1) model again overfits the data, and only slightly outperforms the simple person mean model or the mean mixed model with random intercepts only.

Table 4

The prediction error estimates when applying blocked CV to the COGITO data set. The first row is the average MSE across variables and the other rows the MSE per variable. The standard error of the estimates can be found in parentheses

[6] We do not apply the clusterwise VAR(1) model in the current manuscript because this model is applied to centered data and therefore the results cannot be compared to the ones reported here.

	Overall mean	Person mean	Mean mixed	AR	VAR	Mixed AR	Mixed VAR	Lasso VAR
Average	2.18 (0.03)	1.39 (0.02)	1.38 (0.02)	1.22 (0.02)	1.35 (0.02)	1.20 (0.02)	1.21 (0.02)	1.27 (0.02)
Rumination	2.60 (0.08)	1.92 (0.07)	1.92 (0.07)	1.56 (0.06)	1.78 (0.08)	1.55 (0.06)	1.58 (0.06)	1.65 (0.06)
Guilty	2.65 (0.14)	1.11 (0.06)	1.11 (0.06)	0.82 (0.05)	0.88 (0.05)	0.82 (0.05)	0.82 (0.04)	0.92 (0.06)
Unhappy	1.80 (0.06)	1.30 (0.05)	1.30 (0.05)	1.22 (0.05)	1.29 (0.05)	1.21 (0.05)	1.21 (0.05)	1.19 (0.05)
Down	3.74 (0.12)	2.43 (0.09)	2.42 (0.09)	2.02 (0.08)	2.24 (0.09)	2.00 (0.08)	1.99 (0.08)	2.19 (0.09)
Loss_act	1.95 (0.07)	1.23 (0.05)	1.23 (0.05)	1.15 (0.05)	1.32 (0.05)	1.14 (0.05)	1.16 (0.05)	1.18 (0.05)
Loss_int	1.82 (0.05)	0.93 (0.04)	0.93 (0.04)	0.88 (0.04)	0.95 (0.04)	0.87 (0.04)	0.86 (0.04)	0.88 (0.04)
Sleep_qual	2.09 (0.06)	1.55 (0.06)	1.55 (0.06)	1.52 (0.06)	1.75 (0.08)	1.50 (0.06)	1.52 (0.06)	1.55 (0.06)
Loss_energy	0.83 (0.03)	0.62 (0.02)	0.62 (0.02)	0.56 (0.02)	0.63 (0.03)	0.55 (0.02)	0.56 (0.02)	0.59 (0.02)

Assessment Data

The Assessment data come from a 7-days ESM study focusing on emotion dynamics. The 95 participants (undergraduate students, 62% female) were prompted 10 times a day to fill out a questionnaire. Various papers are already published on these data (Bringmann et al., 2013; Koval, Kuppens, Allen, & Sheeber, 2012; Pe et al., 2013; Pe, Koval, & Kuppens, 2013), but we focus on the report of Bringmann et al. (2016) and limit ourselves to the emotion items, that were rated on a 100-point slider scale. The six included emotion variables are relaxed, happy, dysphoric, anxious, sad, and angry. Bringmann et al. (2016) applied a

mixed VAR(1) model to the data, and used the estimated parameters as input for a network analysis.

The absolute values of the parameter estimates can be found in Figure 6. All estimated models correspond to stationary processes according to eigenvalue analysis. The shrinkage for the mixed and lasso models has the same effect as for the previous data sets. The prediction error estimates and the associated standard errors are presented in Table 5, both averaged across emotions and for each emotion separately. Note that the MSE values are substantially higher compared to the ones of the MindMaastricht and the COGITO data sets because the item scale is larger for these data. As was the case for the previous data set, the mixed AR(1) model and the mixed VAR(1) model have nearly identical, and the lowest, prediction error estimates, with the mixed AR(1) model being preferred based on the one standard error rule. Again, the mixed models are followed by the person-specific AR(1) model, which in turn is followed by the lasso VAR(1) model. The person-specific VAR(1) models are clearly overfitting the data, because the prediction error estimate is higher than when using a person mean model or a mean mixed model. We therefore recommend to not interpret these estimates.

Table 5

The prediction error estimates when applying blocked CV to the Assessment data. The first row is the average MSE across variables and the other rows the MSE per variable. The standard error of the estimates can be found in parentheses

	Overall mean	**Person mean**	**Mean mixed**	**AR**	**VAR**	**Mixed AR**	**Mixed VAR**	**Lasso VAR**

Average	433.99 (4.81)	295.83 (3.58)	295.22 (3.55)	262.90 (3.40)	360.59 (22.51)	257.12 (3.32)	256.05 (3.36)	280.58 (4.76)
Happy	594.39 (10.13)	392.89 (8.74)	392.19 (8.62)	339.83 (8.02)	445.69 (26.14)	332.41 (7.78)	331.74 (7.84)	367.16 (9.78)
Angry	304.77 (11.77)	226.78 (9.31)	226.25 (9.3)	212.59 (9.07)	352.13 (89.24)	207.06 (8.95)	205.88 (9.06)	216.05 (9.25)
Sad	412.52 (12.74)	275.91 (8.80)	275.30 (8.77)	231.34 (7.96)	270.74 (9.29)	225.27 (7.85)	224.12 (7.94)	240.33 (8.20)
Anxious	268.87 (11.07)	175.24 (7.18)	174.98 (7.17)	164.60 (7.10)	191.97 (8.93)	160.69 (7.04)	159.62 (7.12)	167.47 (7.3)
Dysphoric	411.63 (13.64)	223.68 (7.68)	223.40 (7.67)	192.26 (7.19)	216.79 (7.87)	187.56 (7.02)	184.51 (6.95)	198.47 (7.26)
Relaxed	611.75 (10.01)	480.46 (9.80)	479.22 (9.60)	436.76 (9.61)	686.25 (96.64)	429.71 (9.25)	430.42 (9.39)	494.00 (21.04)

Discussion

VAR(1) models are being increasingly applied in psychological research to study within-person dynamics. The recent popularity of the network paradigm has further contributed to the popularity of the VAR(1) model. However, VAR(1) models are complex, because many cross-lagged effects are estimated on top of the autoregressive effects. Yet, to the best of our knowledge, researchers do not verify whether the VAR(1) model is overfitting their data. When overfitting occurs, error specific to the data set at hand is modeled, and as a consequence the estimated model will have difficulties generalizing to unseen data. In such cases, one cannot trust that the estimated parameters indeed characterize the individual from whom the data at hand were sampled (let alone other interpretation difficulties due to differences between the variables in scale and variance (Bulteel et al., 2016b), and to contemporaneous relations between the variables (Bulteel et al., 2016a; Molenaar & Lo,

2016), which is an area of ongoing research). Therefore, the main aim of this study was to quantify the predictive accuracy of the VAR(1) model for current psychological applications, and compare it to other often used models for multivariate time series data of multiple persons.

CV techniques are a versatile tool for assessing the predictive accuracy of stochastic models in general. In case of time series data, both standard CV techniques and variants that take serial dependence into account have been advocated. However, no clear recommendations are available on which approach to adopt. In this paper, we conducted a simulation study with settings mimicking current psychological applications. We showed that approaches modified for time series analysis outperform the standard techniques, however the differences are rather small. In line with Bergmeir and Benítez (2012), we recommend the use of blocked CV. It takes time dependence into account and is easy to use as no additional parameters need to be specified. In addition, it is computationally faster than leave-one-out CV (and APE).

Relying on these simulation results, we applied blocked CV to assess the predictive accuracy of different AR(1) and VAR(1) models, and a number of benchmark models for three state-of-the art psychological multivariate time series data sets. The largest improvement of the prediction error is obtained when accounting for individual differences, as is clear from the difference between results for the overall mean model and the person mean model. Including parameters capturing the time dependence further improved the predictions, indicating that studying time dynamics is worthwhile to pursue. The AR(1) and VAR(1) mixed models had in general the best performance. Following the one standard error rule of Hastie et al. (2009), the mixed AR(1) model is preferred for reasons of parsimony, even though it might be more plausible that the VAR(1) model is the data generating model. While the person-specific AR(1) model also performed well, the person-specific VAR(1) model was

clearly overfitting the data in all three presented applications. The person-specific lasso VAR(1) model only partly resolved these overfitting problems.

In sum, the VAR(1) model (whether the person-specific or mixed variant) did not outperform the less complex AR(1) model for the three prototypical applications considered here. Furthermore, if one expects the data generating model to be even more complex (e.g., non-stationary, non-linear), the burden on the data will be even higher as more complex models are even more prone to overfitting. We do not claim however that more theoretically driven applications of constrained VAR models in which only a few a priori specified lagged effects are estimated would be unfeasible. On the contrary, we strongly believe that if such hypotheses are available they should be incorporated in the analysis, as probably holds for all data-analytical practices. Neither do we want to claim that the VAR(1) model has no value for psychological research. Rather we showed that it is not meaningful to analyze the presented typical applications with a VAR(1) model. The reason might be that the VAR(1) model is inappropriate, or that there is an insufficient number of time points relative to the number of parameters to properly fit a VAR(1) model.

For future data collection, an important question thus pertains to the required number of measurement occasions to properly estimate the VAR(1) coefficients. This number will depend on the specific characteristics of the data set at hand such as the magnitude of the off-diagonal VAR(1) coefficients and the number of variables. Because (person and time) dependent observations contribute less information (compared to independent data), analytical derivations for the required number of observations are not straightforward, and simulations are necessary (e.g., Cools, De Fraine, Van den Noortgate, & Onghena, 2009, in the context of mixed models) to estimate the required number of observations. This manuscript comes with a simulation script that can be used for this purpose, as researchers can enter specific VAR(1)

coefficients and evaluate how both the estimation accuracy and predictive accuracy of the different models that we compared are influenced by the number of observation per person[7].

Ultimately, we would recommend to not only ensure that data are collected at enough measurement occasions to obtain proper estimates, but to design the study in such a way that one can decide whether the AR(1) or the VAR(1) model is more likely to have generated the data. In the context of well-studied retention functions in cognitive psychology, Navarro, Pitt, and Myung (2004) proposed a landscaping analysis to decide between a few models. By examining how well one model fits data simulated with another model and vice versa, the landscape sheds light on the distinguishability of the models, on the potential of the data to discriminate between the models, and on how particular data characteristics influence these capabilities. A complication however is that sufficient knowledge on the psychological process under investigation is required, which is still limited for the dynamical processes studied here.

Inspecting the results, the good performance of the mixed models suggests that for the data sets under study, borrowing strength from other participants in the estimation procedure helps to avoid overfitting and improve the predictive accuracy. This finding contrasts with the recommendations of Gates and Molenaar (2012) that pooling data across persons can give misleading results for heterogeneous samples (which is likely to be the case), but is in line with the findings of Liu (2017) when comparing the parameter accuracy of person-specific and mixed AR models. A possible explanation for our results is the relatively low number of time points per person and the somewhat higher number of variables, but further research is required.

[7] https://www.dropbox.com/s/tolsgza6uvdjm6s/Replication%20Package.zip?dl=0

Another direction for future research is to investigate the performance of other frequently used model selection strategies, for instance the well-known information criteria: Akaike Information Criterion (AIC; Akaike, 1974) and Bayesian Information Criterion (BIC; Schwarz, 1978). These criteria are computationally faster than using CV (but recent computational advances may speed up the CV methods considerably; e.g., Mestdagh, Verdonck, Duisters, & Tuerlinckx, 2015), however applying them is not straightforward because their calculation is based on the number of parameters, which is not clear for a mixed model. Moreover, a mixed model requires modifications of the information criteria (Vaida & Blanchard, 2005).

Finally, we focused on the most popular variants of autoregressive models. However, we suppose that other models may exist that minimize the prediction error even further for the presented applications. Specifically, using variable selection procedures such as lasso, within a mixed model may be a fruitful approach (Müller, Scealy, & Welsh, 2013). Another option is GIMME, a model that estimates both the contemporaneous and the lagged effects while setting some of the coefficients to zero (Gates & Molenaar, 2012). Future research could further compare and propose time series models that are best suited for this kind of psychological data sets.

To conclude, we showed the importance of looking at the predictive accuracy of complex within-person dynamical models in general, and the VAR(1) model specifically. For the presented typical data sets, our analyses suggest that person-specific VAR(1) models are clearly overfitting the data, and that even a mixed variant does not outperform the simpler mixed AR(1) models with regard to predictive accuracy. This is an important finding for psychological practice. Indeed, if the VAR(1) models are overfitting to some extent, it is not meaningful to interpret its parameters or visualize the results in a network (even if VAR(1) models are theoretically more plausible). Of course, the main message is that it is necessary to

assess the predictive accuracy for each future application to avoid making meaning of overfitted models. It depends on the size of the data set and on the complexity of the model whether overfitting will occur.

References

Abegaz, F., & Wit, E. (2013). Sparse time series chain graphical models for reconstructing genetic networks. *Biostatistics, 14*, 586-599. doi: 10.1093/biostatistics/kxt005

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716-723. doi: 10.1109/TAC.1974.1100705

Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine, 66*, 411-421.

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences, 191*, 292-213. doi: 10.1016/j.ins.2011.12.028

Bolger, N., & Laurenceau, J.-P. (2013). *Intensive longitudinal methods : An introduction to diary and experience sampling research*. New York, NY: The Guilford Press.

Borsboom, D., & Cramer, A. O. J. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology, 9*, 91-121. doi: 10.1146/annurev-clinpsy-050212-185608

Bos, E. H., Hoenders, R., & de Jonge, P. (2012). Wind direction and mental health: A time-series analysis of weather influences in a patient with anxiety disorder. *BMJ Case Reports*. doi: 10.1136/bcr-2012-006300

Brandt, P. T., & Williams, J. T. (2007). *Multiple time series models*. Thousand Oaks, CA: Sage Publications.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science, 16*, 199-231. doi: 10.1214/ss/1009213726

Bringmann, L. F., Pe, M. L., Vissers, N., Ceulemans, E., Borsboom, D., Vanpaemel, W., Tuerlinckx, F., & Kuppens, P. (2016). Assessing temporal emotion dynamics using networks. *Assessment, 23*, 425-435. doi:10.1177/1073191116645909

Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., Borsboom, D., & Tuerlinckx, F. (2013). A network approach to psychopathology: New insights into clinical longitudinal data. *PLoS ONE, 8*, e60188, 1-13. doi:10.1371/journal.pone.0060188

Bulteel, K., Tuerlinckx, F., Brose, A., & Ceulemans, E. (2016a). Clustering vector autoregressive models: Capturing qualitative differences in within-person dynamics. *Frontiers in Psychology, 7*, 1540. doi:10.3389/fpsyg.2016.01540

Bulteel, K., Tuerlinckx, F., Brose, A., & Ceulemans, E. (2016b). Using raw VAR regression coefficients to build networks can be misleading. *Multivariate Behavioral Research, 51*, 330-344. doi:10.1080/00273171.2016.1150151

Burman, P., Chow, E., & Nolan, D. (1994). A cross-validatory method for dependent data. *Biometrika, 81*, 351-358. doi: 10.2307/2336965

Coifman, K. G., Bonanno, G. A., & Rafaeli, E. (2007). Affect dynamics, bereavement, and resilience to loss. *Journal of Happiness Studies, 8*, 371-392. doi: 10.1007/s10902-006-9014-5

Cools, W., De Fraine, B., Van den Noortgate, W., & Onghena, P. (2009). Multilevel design efficiency in educational effectiveness research. *School Effectiveness and School Improvement, 20*, 357-373. doi: 10.1080/09243450902850176

Douglas, H. E. (2009). Reintroducing prediction to explanation. *Philosophy of Science, 76*, 444-463. doi: 10.1086/648111

Gates, K. M., & Molenaar, P. C. M. (2012). Group search algorithm recovers effective connectivity maps for individuals in homogeneous and heterogeneous samples. *NeuroImage, 63*, 310-319. doi: 10.1016/j.neuroimage.2012.06.026

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association, 70*, 320-328. doi: 10.2307/2285815

Geschwind, N., Peeters, F., Drukker, M., van Os, J., & Wichers, M. (2011). Mindfulness training increases momentary positive emotions and reward experience in adults vulnerable to depression: A randomized controlled trial. *Journal of Consulting and Clinical Psychology, 79*, 618-628. doi: 10.1037/a002459

Hamaker, E. L., Ceulemans, E., Grasman, R. P. P. P., & Tuerlinckx, F. (2015). Modeling affect dynamics: State-of-the-art and future challenges. *Emotion Review, 7*, 316-322. doi:10.1177/1754073915590619

Hamilton, J. D. (1994). *Time series analysis*. Princeton, NJ: Princeton University Press.

Harrison, L., Penny, W. D., & Friston, K. (2003). Multivariate autoregressive modeling of fMRI time series. *NeuroImage, 19*, 1477-1491. doi: 10.1016/S1053-8119(03)00160-5

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY: Springer.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. New York, NY: Routledge.

Hsu, N.-J., Hung, H.-L., & Chang, Y.-M. (2008). Subset selection for vector autoregressive processes using Lasso. *Computational Statistics & Data Analysis, 52*, 3645-3657. doi: 10.1016/j.csda.2007.12.004

Koval, P., Kuppens, P., Allen, N. B., & Sheeber, L. (2012). Getting stuck in depression: The roles of rumination and emotional inertia. *Cognition and Emotion, 26*, 1412-1427. doi: 10.1080/02699931.2012.667392

Liu, S. (2017). Person-specific versus multilevel autoregressive models: Accuracy in parameter estimates at the population and individual levels. *British Journal of Mathematical and Statistical Psychology*. Advance online publication. doi: 10.1111/bmsp.12096

Lodewyckx, T., Tuerlinckx, F., Kuppens, P., Allen, N. B., & Sheeber, L. B. (2011). A hierarchical state space approach to affective dynamics. *Journal of Mathematical Psychology, 55,* 68-83. doi:10.1016/j.jmp.2010.08.004

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Berlin, Germany: Springer.

Mestdagh, M., Verdonck, S., Duisters, K., & Tuerlinckx, F. (2015). Fingerprint resampling: A generic method for efficient resampling. *Scientific Reports, 5,* 16970, 1-21. doi:10.1038/srep16970

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary research and perspectives, 2,* 201-218. doi: 10.1207/s15366359mea0204_1

Molenaar, P. C. M., & Campbell, C. G. (2009). The new person-specific paradigm in Psychology. *Current Directions in Psychological Science, 18,* 112-117. doi: 10.1111/j.1467-8721.2009.01619.x

Molenaar, P. C. M., & Lo, L. L. (2016). Alternative forms of Granger causality, heterogeneity, and nonstationarity. In W. Wiedermann & A. von Eye (Eds.), *Statistics and causality: Methods for applied empirical research* (pp. 205-230). doi: 10.1002/9781118947074.ch9

Müller, S., Scealy, J. L., & Welsh, A. H. (2013). Model selection in linear mixed models. *Statistical Science, 28,* 135-167. doi: 10.1214/12-STS410

Narravo, D. J., Pitt, M. A., & Myung, I. J. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology, 49,* 47-84. doi: 10.1016/j.cogpsych.2003.11.001

Pe, M. L., Kircanski, K., Thompson, R. J., Bringmann, L. F., Tuerlinckx, F., Mestdagh, M., … Gotlib, I. H. (2015). Emotion-network density in major depressive disorder. *Clinical Psychological Science, 3,* 292-300. doi: 10.1177/2167702614540645

Pe, M. L., Koval, P., & Kuppens, P. (2013). Executive well-being: Updating of positive stimuli in working memory is associated with subjective well-being. *Cognition, 126*, 335-340. doi: 10.1016/j.cognition.2012.10.002

Pe, M. L., Raes, F., Koval, P., Brans, K., Verduyn, P., & Kuppens, P. (2013). Interference resolution moderates the impact of rumination and reappraisal on affective experiences in daily life. *Cognition and Emotion, 27*, 492-501. doi: 10.1080/02699931.2012.719489

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences, 6*, 421-425. doi: 10.1016/S1364-6613(02)01964-2

Racine, J. (2000). Consistent cross-validatory model-selection for dependent data: *hv*-block cross-validation. *Journal of Econometrics, 99*, 39-61. doi: 10.1016/S0304-4076(00)00030-0

Reich, J. W., Zautra, A. J., & Davis, M. (2003). Dimensions of affect relationships: Models and their integrative implications. *Review of General Psychology, 7*, 66-83. doi: 10.1037/1089-2680.7.1.66

Rissanen, J. (1986). Order estimation by accumulated prediction errors. *Journal of Applied Probability, 23*, 55-61. doi: 10.2307/3214342

Roebroeck, A., Formisano, E., & Goebel, R. (2005). Mapping directed influence over the brain using Granger causality and fMRI. *NeuroImage, 25*, 230-242. doi: 10.1016/j.neuroimage.2004.11.017

Rosmalen, J. G. M., Wenting, A. M. G., Roest, A. M., de Jonge, P., & Bos, E. H. (2012). Revealing causal heterogeneity using time series analysis of ambulatory assessments: Application to the association between depression and physical activity after myocardial infarction. *Psychosomatic Medicine, 74*, 377-386. doi: 10.1097/PSY.0b013e3182545d47

Rothman, A. J., Levina, E., & Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics, 19*, 947-962. doi: 10.1198/jcgs.2010.09188

Sbarra, D. A., & Allen, J. J. B. (2009). Decomposing depressing: On the prospective and reciprocal dynamics of mood and sleep disturbances. *Journal of Abnormal Psychology, 118*, 171-182. doi: 10.1037/a0014375

Schmiedek, F., Lövdén, M., & Lindenberger, U. (2010). Hundred days of cognitive training enhance broad cognitive abilities in adulthood: Findings from the COGITO study. *Frontiers in Aging Neuroscience, 2*, 27. doi: 10.3389/fnagi.2010.00027

Schmitz, B., & Skinner, E. (1993). Perceived control, effort, and academic performance: interindividual, intrainindividual, and multivariate time-series analysis. *Journal of Personality and Social Psychology, 64*, 1010-1028. doi: 10.1037/0022-3514.64.6.1010

Schuurman, N.K., Ferrer, E., de Boer-Sonnenschein, M., & Hamaker, E.L. (2016). How to compare cross-lagged associations in a multilevel autoregressive model. *Psychological Methods, 21*, 206-221. doi: 10.1037/met0000062

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461-464. doi: 10.1214/aos/1176344136

Shmueli, G. (2010). To explain or to predict? *Statistical Science, 25*, 289-310. doi: 10.1214/10-STS330

Snijders, T. A. B. (1988). On cross-validation for predictor evaluation in time series. In T.K. Dijkstra (Ed.), *On model uncertainty and its statistical implications* (pp. 56-69). New York, NY: Springer-Verlag.

Snippe, E., Bos, E. H., van der Ploeg, K. M., Sanderman, R., Fleer, J., & Schroevers, M. J. (2015). Time-series analysis of daily changes in mindfulness, repetitive thinking, and depressive symptoms during mindfulness-based treatment. *Mindfulness, 6*, 1053-1062. doi: 10.1007/s12671-014-0354-7

Stone, M. (1974). Cross-validation and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological), 36*, 111-147.

Tibshirani, R. (1996). Shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological), 58*, 267-288.

Vaida, F., & Blanchard, S. (2005). Conditional Akaike information for mixed-effects model. *Biometrika, 92*, 351-370. doi: 10.1093/biomet/92.2.351

van der Krieke, L., Emerencia, A. C., Bos, E. H., Rosmalen, J. G. M., Riese, H., Aiello, M., Sytema, S., & de Jonghe, P. (2015). Ecological momentary assessments and automated time series analysis to promote tailored health care: A proof-of-principle study. *JMIR Research Protocols, 4*, e100. doi: 10.2196/resprot.4000

van Gils, A., Burton, C., Bos, E. H., Janssens, K. A. M., Schoevers, R. A., & Rosmalen, J. G. M. (2014). Individual variation in temporal relationships between stress and functional somatic symptoms. *Journal of Psychosomatic Research, 77*, 34-39. doi: 10.1016/j.jpsychores.2014.04.006

Wichers, M. (2014). The dynamic nature of depression: A new micro-level perspective of mental disorder that meets current challenges. *Psychological Medicine, 44*, 1349-1360. doi: 10.1017/S0033291713001979

Wild, B., Eichler, M., Friederich, H.-C., Hartmann, M., Zipfel, S., & Herzog, W. (2010). A graphical vector autoregressive modelling approach to the analysis of electronic diary data. *BMC Medical Research Methodology, 10*, 28. doi: 10.1186/1471-2288-10-28

Zautra, A. J., Berkhof, J., & Nicolson, N. A. (2002). Changes in affect interrelations as a function of stressful events. *Cognition and Emotion, 16*, 309-318. doi: 10.1080/02699930143000257

Zautra, A. J., Reich, J. W., Davis, M. C., Potter, P. T., & Nicolson, N. A. (2000). The role of stressful events in the relationship between positive and negative affects: Evidence from field and experimental studies. *Journal of Personality, 68*, 927-951. doi: 10.1111/1467-6494.00121

Zheng, Y., Wiebe, R. P., Cleveland, H. H., & Molenaar, P. C. M., Harris, K. S. (2013). An idiographic examination of day-to-day patterns of substance use craving, negative affect, and tobacco use

among young adults in recovery. *Multivariate Behavioral Research, 48*, 241-266. doi: 10.1080/00273171.2013.763012